Algoritmo Genético Multi-Objetivo Para Seleção de Variáveis em Problemas de Classificação Multivariada: Um Estudo de Caso na Verificação de Adulteração de Biodiesel

Lucas de Almeida Ribeiro
Anderson da Silva Soares
Instituto de Informática
Universidade Federal de Goiás
Goiânia, Goiás
Email: (lucasaribeiro, anderson)@inf.ufg.br

Clarimar José Coelho
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Goiás
Goiânia, Goiás

Fabrízzio A. A. Melo Nunes Soares
Telma Woerle de Lima
Carlos Antônio Campos Jorge
Instituto de Informática
Universidade Federal de Goiás
Goiânia, Goiás

***Resumo*—Este trabalho tem por objetivo propor uma formulação multiobjetivo utilizando algoritmos genéticos para o problema de seleção de variáveis em calibração multivariada. Apresenta-se um problema envolvendo a classificação de amostras de biodiesel para detecção de adulteração por meio de um classificador denominado análise discriminante linear. O papel do algoritmo genético multiobjetivo consiste em reduzir a dimensionalidade do conjunto original de variáveis de modo a proporcionar um modelo de classificação mais robusto e consequentemente de melhor capacidade de generalização. Em particular o trabalho faz uma implementação da versão de elitismo rápido do algoritmo genético de ordenação por não dominância (NSGA-II). É feita uma comparação entre uma implementação monoobjetivo e multiobjetivo com relação ao modelo e robustez em relação à presença de ruído. Os resultados mostram que em média o algoritmo genético monoobjetivo seleciona 20 variáveis e tem uma taxa de erro de 14% e o multiobjetivo seleciona 7 variáveis e tem uma taxa de erro de 11%. Foi possível demonstrar que a formulação multiobjetivo proporciona modelos de classificação com menor sensibilidade a ruído instrumental quando comparada a formulação monoobjetivo.**

I. Introduction

A quimiometria pode ser definida como a aplicação de métodos matemáticos e estatísticos em dados de natureza química [9]. Nesse contexto uma possível aplicação de métodos quimiométricos consiste na determinação ou classificação de elementos de interesse em um determinado composto, fazendo uso de medidas indiretas obtidas por análise instrumental de sinal.

Uma das características, dos métodos instrumentais atuais, é a possibilidade de gerar inúmeras variáveis a partir de uma única amostra[14]. Um exemplo é a intensidade de absorção em até milhares de comprimentos de onda, que podem ser registrados em um único espectro [13]. Os espectros são o resultado da aplicação de energia na matéria, onde espectroscopia é o estudo dessa interação. Em relação à energia, esta pode ser absorvida pela matéria, e a quantidade de absorção é dependente do tipo do composto[19]. Porém nem todas as variáveis medidas estão associadas ao composto de interesse. A associação entre variáveis medidas e compostos pode ser feita a partir de técnicas estatísticas multivariadas tais como a análise discriminante[23].

A análise discriminante é uma técnica estatística para a discriminação e classificação de elementos em grupos conhecidos [3]. Assim dadas as variáveis de uma amostra, pode-se aloca-la em sua classe correspondente através de uma regra de associação.[12]

A análise discrimintante pode ser entendida como um modelo de reconhecimento de padrões com enfoque estatístico que usa um subconjunto de observações multivariadas, onde sua classificação é conhecida, chamado de grupo de treinamento. A partir desse grupo obtém uma regra de associação ou uma função classificadora, para a coleção inteira [6]. Um método bem conhecido para se fazer a classificação é a Análise Discriminante Linear (LDA). Esse método utiliza-se da inversa da matriz de covariância conjunta das classes, para fazer a classificação dos elementos [24]. Um dos problemas da aplicação da análise discriminante linear remete a problemas inerentes à inversão de matrizes, como o de colinearidade ou quase colinearidade das colunas da matriz de covariâncias utilizada [16], [17].

Uma forma de minimizar as variáveis colineares é removê-las do modelo por meio de técnicas de seleção de variáveis. Diversos algoritmos têm sido utilizados para realizar esse processo, dentre eles destacam-se os algoritmos genéticos [11][20], o algoritmo das projeções sucessivas[21] e o algoritmo *stepwise* [20].

Algoritmos genéticos são utilizados para a seleção de variáveis, obtidas por meio da espectroscopia de Raman, com o objetivo de realizar a classificação de madeira. O algoritmo genético foi eficiente nesse trabalho pro selecionar as faixas que maximizavam as diferenças das classes [11]. Algoritmos genéticos, algoritmo stepwise, e o algoritmo de projeções sucessivas são utilizados para a seleção de variáveis na classificação de canetas de tintas azul em tipos e marcas. O método desenvolvido, com os diferentes processos de seleção de variáveis, foi hábil para diferenciar as marcas e tipos de canetas [20]. Em contrapartida aos resultados satisfatórios,

estas implementações utilizam-se de abordagens com objetivos únicos. Assim para se maximizar o ganho em relação ao objetivo pretendido o algoritmo procura incluir um número cada vez maior de variáveis, sem considerar a possibilidade de modelos com um menor número e poder de discriminação semelhante.

Neste trabalho é proposto o uso de uma formulação multiobjetivo de algoritmo de seleção de variáveis para análise discriminante linear em problemas de classificação multivariada. Em particular, é proposto o uso do algoritmo genético de ordenação por não dominância (NSGA-II) para seleção de variáveis com dois objetivos: minimização do erro de classificação e número de variáveis utilizadas. Espera-se que ao incluir um segundo objetivo, além do erro de classificação, o modelo final possua melhor capacidade de generalização, quando comparado a abordagem tradicional de monoobjetivo. Como estudo de caso, utiliza-se um problema do mundo real envolvendo a classificação de misturas de combustíveis de diesel/biodiesel para verificação de adulteração.

II. REVISÃO TEÓRICA

A. ANÁLISE DISCRIMINANTE

Análise Discriminante é um tipo de técnica supervisionada de reconhecimento de padrões, onde a densidade de probabilidades dos objetos pertencentes a classe são modeladas como distribuições normais multivariadas. Assumindo a mesma probabilidade, a priori, para todas as classes consideradas, a classificação é dada calculando-se a probabilidade em cada uma das classes possíveis, a classe que obter a maior probabilidade é assumida como a classe a que o elemento pertence[2]. Para se fazer a classificação utiliza-se da inversa da matriz de covariância das classes. Ao assumir a mesma matriz de covariância, Σ, para todas as classes, a região de separação é dada pela equação1 [10].

$$(x-\mu_{j_1})^T\Sigma^{-1}(x-\mu_{j_1}) = (x-\mu_{j_2})^T\Sigma^{-1}(x-\mu_{j_2}) \quad (1)$$

E assim o termo geométrico $(x)^T\Sigma^{-1}(x)$ se torna independente da classe formando uma superfície linear de decisão em hiperplanos no $\Re^k$. E o processo de classificação fica associado ao conceito da distância de Mahalanobis, definida por r da equação 2 [10].

$$r^2(x,\mu_j)^T = (x-\mu_j)^T\Sigma^{-1}(x-\mu_j) \quad (2)$$

Essa distância é constante na classe j e μ_j é o vetor média para a classe j, observando novamente que Σ é único para todas as classes e pode ser estimado a partir das matrizes de covariância das classes. O cálculo é feito a partir da média ponderada das matrizes de covariância individuais de cada classe [10].

Assim, para fazer a classificação do elemento x, é suficiente fazer o cálculo da distância de Mahalanobis, em cada uma das classes, e classificar x na que obtiver o menor valor. Caso seja verificada a homocedasticidade, o LDA será ótimo para classificação, no sentido de minimização de erros [10].

Vale notar que o cálculo da distância de Mahalanobis depende da estabilidade do cálculo da inversa da matriz de covariâncias. Para tanto, é necessário que as variáveis consideradas não possuam colinearidade entre si, caso contrário o cálculo da inversa não será possível ou poderá haver problemas de mal condicionamento da matriz inversa. Neste sentido, um algoritmo de seleção de variáveis como o algoritmo genético poderá ser modelado para se selecionar variáveis representativas do problema e com baixa colinearidade entre si [10].

B. ALGORITMO GENÉTICO

Algoritmos genéticos (GA) são uma técnica de busca inspirada em uma metáfora do processo biológico de evolução natural, onde há preferência pelo indivíduo mais adaptado [15].Os algoritmos genéticos também podem ser definidos como técnicas heurísticas de otimização global, onde diferente de métodos que seguem a derivadas de função para encontrar ótimos, podendo ficar restrito a ótimos locais, fazem uma busca global no espaço de soluções [1].

Um algoritmo genético é projetado presumindo que a solução do problema pode ser modelada por um conjunto de parâmetros, muitas vezes em notação binária. Estes parâmetros são organizados como genes de um cromossomo, e a partir dos genes codificados aplica-se uma função de conversão de gene à aptidão. A função de aptidão atribui a característica fenotípica de uma configuração de genes. Indivíduos de uma população com os melhores valores de aptidão (melhores características fenotípicas) devem ser priorizados na produção de uma nova população [7].

O funcionamento do algoritmo genético é dado pelo processo: calcule o valor de aptidão da população atual, escolha os indivíduos que produzirão a nova população, produza a partir de operadores genéticos sobre os indivíduos escolhidos uma nova população. O processo se repete por um número fixo de gerações, ou até a obtenção de indivíduo(s) com a característica desejada[15].

A função de aptidão é projetada de acordo com o problema a ser tratado. Uma população pode derivar outra por meio de dois processos: cruzamento e mutação. No cruzamento partes do cromossomo de um dos indivíduos selecionados é agrupada com partes do cromossomo de outro indivíduo para a formação de um novo. A mutação é o processo de mudança de valor de um dos genes que constituem o cromossomo [15]. Esses algoritmos tentam a cada geração proporcionar soluções melhores que as anteriores.

C. ALGORITMO GENÉTICO MULTI-OBJETIVO

Diversos problemas do mundo real necessitam a otimização simultânea de múltiplos objetivos [8]. No caso de problemas com único objetivo, os algoritmos utilizados para otimização buscam encontrar a melhor solução do problema. Porém, quando se trata de problemas multiobjetivos, pode não haver uma solução ótima, única, para todos os objetivos. Neste contexto, em problemas de otimização multiobjetivo existe um subconjunto no espaço de soluções que é superior às demais soluções, em todos os objetivos analisados. Esse subconjunto é conhecido como soluções ótimas de Pareto ou soluções não dominadas [2]. A escolha de uma única solução na coleção de soluções ótimas de Pareto depende do conhecimento dos fatores relacionados ao problema. Uma solução escolhida em determinado modelo, pode não ser a melhor em outro modelo,

ou ambiente. Desta forma, em análises multiobjetivos, deve-se fazer o uso de medidas alternativas na escolha da solução dentro da coleção de soluções ótimas de Pareto [22].

1) NSGA-II: O NSGA implementa o conceito de não dominância, classificando a população em fronteiras de acordo com seu nível de dominância. Em cada geração as melhores soluções são alocadas na primeira fronteira e as piores são alocadas na última. O processo de alocação termina assim que todos os indivíduos estão alocados, em suas respectivas fronteiras. Após acontecer o processo de alocação, os indivíduos da primeira fronteira não são dominados por nenhum outro indivíduo, mas dominam os da segunda fronteira, e os indivíduos da i-ésima fronteira dominam os indivíduos da (i+1)-ésima fronteira. Após vários anos de críticas em relação ao NSGA referentes a problemas como o de alta complexidade computacional e demora na realização do processo de elitismo, foi proposto uma nova versão do NSGA [5]. A principal diferença encontrada nessa versão, o NSGA-II, é a forma como o operador de seleção é aplicado, operador este que é dividido em dois processos, figura 1: ordenação não dominada rápida(ONDR), e ordenaço pela distancia *crowding*, que buscam simplificar o processo de comparação dos indivíduos, e diminuir a similaridade de soluções, respectivamente.

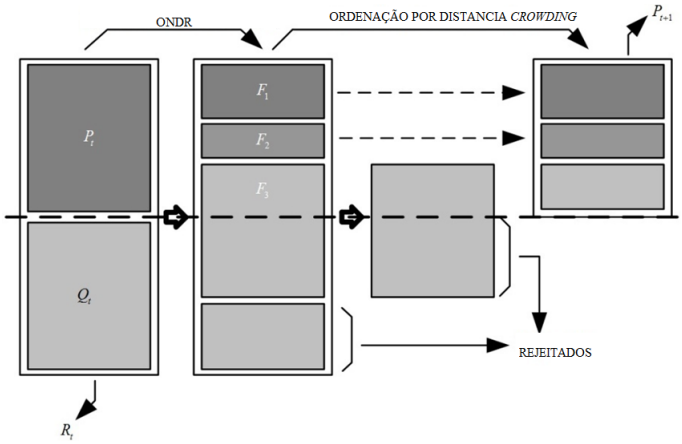


Figura 1. Processo de Ordenação Pela Não Dominância e Por Distância *crowding*

A figura 1 apresenta como os dois métodos fazem a ordenação dos indivíduos para a escolha dos pais, onde na ordenação não dominada rápida os indivíduos são classificados em fronteiras(F_1, F_2, F_3) e os indivíduos são colocados em ordem de acordo com a fronteira. Já a ordenação pela distância *crowding* busca privilegiar os indivíduos com menor similaridade entre si, para manter uma população mais heterogênea. Os indivíduos que estão acima do limite de pais de uma geração são rejeitados.

III. MATERIAS E MÉTODOS

A. Conjunto de Dados

O conjunto de dados utilizado neste trabalho foi obtido por meio de espectroscopia no Infravermelho Próximo(NIR - *Near InfraRed*). A classificação de dados de calibração multivariada utilizando-se de LDA não é nova na literatura. E o conjunto de dados já foi trabalhado em [18] e [4], utilizando-se de outras técnicas de seleção de variáveis.

Os dados foram obtidos a partir de amostras de misturas de biodiesel. E estão distribuídos em quatro classes distintas: diesel livre de biodiesel e óleo vegetal cru (D), misturas contendo diesel, biodiesel e óleo vegetal cru (OBD), misturas de diesel e óleo vegetal cru (OD), e por último misturas contendo diesel e 5% de biodiesel como previsto por lei(B5)[18].

A espectrometria no infravermelho próximo (NIR) foi utilizada pois o padrão europeu para a avaliação de qualidade de biodiesel $EM14078$ considera a absorção de carbonila na região infravermelho médio (MIR - *Middle InfraRed*) em um método não variável em 1745 cm-1, mesma região de absorção dos grupos de carbono presentes nos óleos vegetais crus, assim uma adulteração com óleos vegetais pode não ser detectada a partir do padrão europeu, e por isso vários estudos buscam outras regiões de absorção para a detecção, muitas no MIR e no NIR. Os problemas derivados do uso de combustíveis adulterados por óleos crus estão ligados a problemas no motor como: depósitos de carbono, bloqueio de injeção, incompleta combustão e formação de goma que é uma característica da oxidação e polimerização[18] .

O conjunto de dados é composto de 76 amostras para treinamento, 32 amostras de validação e 32 amostras para teste de classificação. Cada amostra possui 1296 variáveis, que são os comprimentos de onda analisados. Os dados foram medidos, por meio de NIR, na região: 1.0 mm $(8814$ - $3799)$ cm^{-1}[18].

Para a classe D tem-se 19 amostras para treinamento, 8 para validação e 8 para testes. Na classe OBD tem-se 20 amostras para treinamento, 9 para validação e 9 para testes. Na classe OD, 19 amostras para treinamento, 8 para validação e 8 para testes. E para a classe B5 18 amostras para treinamento, 7 para validação e 7 para testes. As misturas foram preparadas usando diferentes óleos vegetais, gordura animal e seus respectivos ésteres. O biodiesel puro e as amostras de óleo foram adquiridos no mercado, de origem de diversas indústrias, todos seguindo as normas brasileiras[18] .

B. MONO GA LDA

A implementação monoobjetivo do algoritmo genético de seleção de variáveis foi executada com a notação binária para a escolha de variáveis. Em que cada cromossomo se constituía das 1296 variáveis, das quais no máximo 32 variáveis eram escolhidas inicialmente, por ser este o número de amostras do conjunto de validação. O elitismo foi escolhido como o método de escolha dos pais para a próxima geração, privilegiando-se através de ordenação os indivíduos da geração mais próxima.

A função de avaliação escolhida foi o número de erros obtidos na classificação do conjunto de validação. O número de gerações escolhido foi de 100, a probabilidade de mutação foi de 5% para cada gene, e o número de indivíduos de cada população ficou limitado à 80.

C. MULTI GA LDA

O cromossomo foi modelado da mesma forma que na abordagem monoobjetivo. As funções de avaliação escolhidas para a modelagem multiobjetivo foram a minimização de erros e a minimização de número de variáveis. Foi estabelecido o mesmo número de gerações e o mesmo tamanho da população em relação ao monoobjetivo.

Algorithm 1 MONO-GA-LDA

Seja P uma população inicial gerada randomicamente;
Seja $conf$ as configurações de mutação dadas;
Seja f o número de pais
Seja gen o número de gerações dadas;
for all $p \in P$ **do**
  $p_f \leftarrow$ funcaoFitness(p);
end for
ORDENAR(P);
$P_p \leftarrow ELITISMO(P, f)$;
for $i \leftarrow 1 to gen$ **do**
  $P_s \leftarrow OPERADORES - GENETICOS(P_p, conf)$;
  for all $p \in P_p$ **do**
    $p_f 1 \leftarrow$ funcaoFitness(p);
    $P \leftarrow P \cup B$;
    ORDENAR(P);
    $P_p \leftarrow ELITISMO(P, f)$;
  end for
end for

A probabilidade de mutação estabelecida foi de em média 2,5%, 50% para o indivíduo e 5% para o gene. O cruzamento entre os pais para se gerar a nova geração foi dado por crossover uniforme. A seleção dos pais para uma nova geração foi feito através do torneio binário. O algoritmo MULTI-GA-LDA está apresentado no Algoritmo 2.

Algorithm 2 MULTI-GA-LDA

Seja P uma população inicial gerada randomicamente;
Seja $conf$ as configurações de mutação dadas;
Seja f o número de pais
Seja gen o número de gerações dadas;
for all $p \in P$ **do**
  $p_f 1 \leftarrow$ funcaoFitness1(p);
  $p_f 2 \leftarrow$ funcaoFitness2(p);
end for
ORDENAR-POR-NAO-DOMINANCIA(P);
$P_p \leftarrow TORNEIO - BINARIO(P)$;
for $i \leftarrow 1 to gen$ **do**
  $P_s \leftarrow OPERADORES - GENETICOS(P_p, conf)$;
  for all $p \in P_p$ **do**
    $p_f 1 \leftarrow$ funcaoFitness1(p);
    $p_f 2 \leftarrow$ funcaoFitness2(p);
    $P \leftarrow P \cup B$;
    ORDENAR-POR-NAO-DOMINANCIA(P);
    $P_p \leftarrow TORNEIO - BINARIO(P)$;
  end for
end for

IV. Resultados e Discussões

O extrato abaixo, tabela I, apresenta os resultados obtidos a partir de 40 execuções do algoritmo MONO-GA-LDA. Apresenta-se o número de variáveis selecionadas e a quantidade de erros obtidas pelo algoritmo MONO-GA-LDA.

O extrato abaixo, tabela II, apresenta os dados obtidos de 40 execuções do algoritmo MULTI-GA-LDA. Apresenta-se o número de variáveis selecionadas e a quantidade de erros obtidas pelo algoritmo MULTI-GA-LDA. Quando comparados os resultados dos dois algoritmos (MONO-GA-LDA e MULTI-GA-LDA) é possível notar que o número de erros é similar. Entretanto, a quantidade de variáveis empregada no classificador é substancialmente menor (na ordem de 3 vezes, em média) quando utilizada a formulação multiobjetivo proposta.

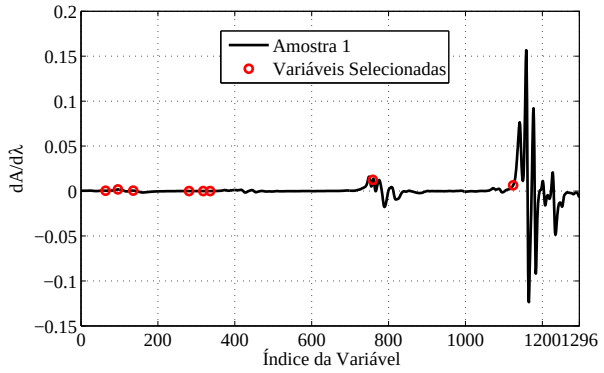
	Maior	Menor	Média	Desvio Padrão
No Var	32	6	19.85	6.72
No Erros	11	0	4.5	1.9

Tabela I. Extrato de dados de execução do Algoritmo MONO-GA-LDA

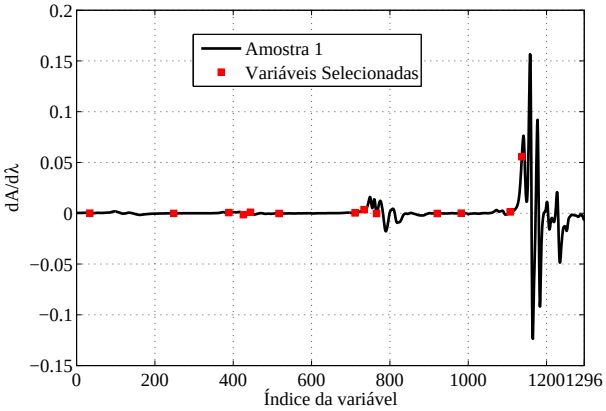
	Maior	Menor	Média	Desvio Padrão
No Var	17	4	6.82	3.19
No Erros	10	0	3.65	2.31

Tabela II. Extrato de dados de execução do Algoritmo MULTI-GA-LDA

Além dos valores apresentados nas tabelas acima foi analisado o cromossomo com melhor taxa de erro obtido pelo algoritmo MULTI-GA-LDA e o melhor cromossomo obtido pelo MONO-GA-LDA. Tais cromossomos serão utilizados deste ponto em diante, e as variáveis que compõem cada um dos cromossomos são apresentadas na figura 2.



(a) MULTI-GA-LDA

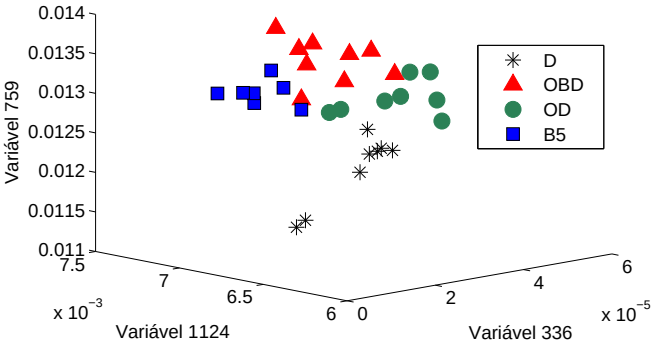


(b) MONO-GA-LDA

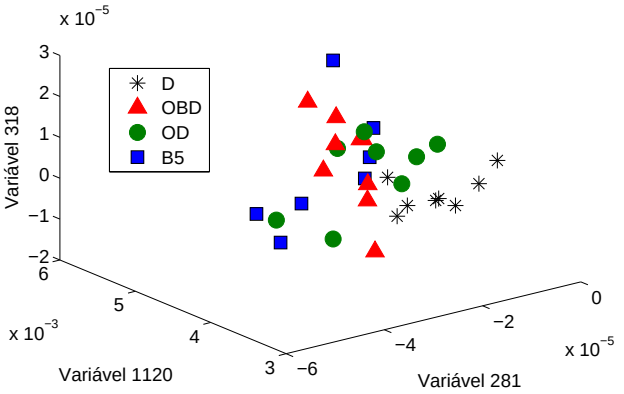
Figura 2. Variáveis que compõem os cromossomos com melhor taxa de erro obtida na classificação das amostras do conjunto de teste de classificação

A figura 3 apresenta a dispersão das amostras do conjunto de teste de classificação, utilizando as duas variáveis de maior discriminabilidade medida pelo coeficiente de Fisher. Nota-

se que utilizando as variáveis selecionadas pelo MULTI-GA-LDA (figura 3(a)) a separação entre as amostras torna-se mais definida quando comparada a dispersão das amostras utilizando as variáveis selecionadas pelo MONO-GA-LDA (figura 3(b)), sugerindo que o classificador será mais eficiente em novas amostras.



(a) MULTI-GA-LDA



(b) MONO-GA-LDA

Figura 3. Dispersão das amostras utilizando as duas variáveis de maior discriminabilidade selecionadas pelos algoritmos MULTI-GA-LDA e MONO-GA-LDA

Utilizando ainda os dois melhores cromossomos obtidos em cada uma das abordagens, foi realizado um estudo de sensibilidade à ruído de medida instrumental. Tal estudo consiste em contaminar as variáveis independentes com um ruído aleatório branco. A figura 4 apresenta a relação entre o número de erros do classificador a medida em que a taxa de ruído presente no sinal é incrementada. O eixo Y apresenta o número de erros apresentados, em média de 50 execuções do classificador em razão do erro ser aleatório, no ambiente com contaminação artificial. E o eixo X apresenta a taxa máxima de ruído, que foram gerados aleatoriamente com valores em módulo, de no máximo a porcentagem apresentada. Como é possível observar, o modelo obtido pelo algoritmo proposto (MULTI-GA-LDA) possui uma menor sensibilidade à presença de ruído instrumental, quando comparado à formulação monoobjetivo. Como instigado, a formulação multiobjetivo leva à modelos mais robustos não apenas em termos de números de variáveis mas também em termos de tolerância a ruídos.

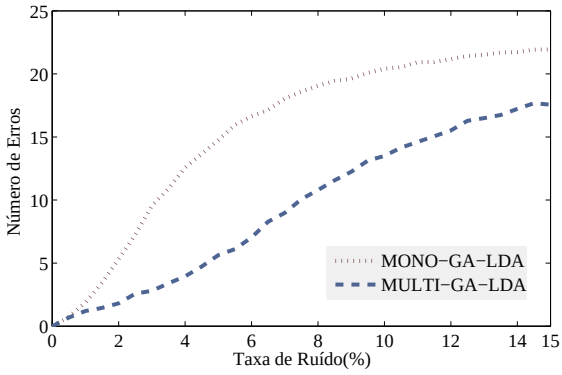


Figura 4. Taxa de Erros em Função Ruído

V. Conclusão

Este trabalho propôs o uso de uma formulação multiobjetivo no problema de seleção de variáveis em análise discriminante linear. Em particular foi proposto o uso do algoritmo NSGA-II para minimizar de forma simultânea o número de erros de classificação e o número de variáveis empregadas para a construção do classificador. Como estudo de caso utilizou-se o problema de verificação de adulteração de biodiesel a partir de dados obtidos por espectofotometro. Os resultados demostraram que a formulação multiobjetivo apresentou desempenho similar a formulação monoobjetivo porém com uma quantidade de variáveis substancialmente menor. Adicionalmente foi feito um estudo de sensibilidade à ruído utilizando-se dos cromossomos com melhor taxa de erros na coleção de teste de classificação, obtidos em cada uma das implementações. A partir do resultado foi possível observar que o modelo de classificação obtido a partir da formulação multiobjetivo possui maior robustez quanto à presença de ruído instrumental.

Agradecimentos

Os autores agradecem à CAPES e à FAPEG (Fundação de Amparo á Pesquisa do Estado de Goiás) pelo apoio financeiro à este projeto.

Referências

[1] Pratibha Bajpai and Manoj Kumar, *Genetic algorithm-an approach to solve global optimization problems*, Indian Journal of computer science AND engineering **1** (2010), no. 3, 199–206.

[2] V Chankong and YY Haimes, *Multi-objective optimization: Pareto optimality*, Concise Encyclopedia of Environmental Systems. Pergamon Press, UK (1993), 387–396.

[3] D. Coomans, D.L. Massart, and L. Kaufman, *Optimization by statistical linear discriminant analysis in analytical chemistry*, Analytica Chimica Acta **112** (1979), no. 2, 97 – 122.

[4] Fernanda Vera Cruz de Vasconcelos et al., *Using near-infrared overtone regions to determine biodiesel content and adulteration of diesel/biodiesel blends with vegetable oils*, Analytica chimica acta **716** (2012), 101–107.

[5] Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and Tanaka Meyarivan, *A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii*, Lecture notes in computer science **1917** (2000), 849–858.

[6] Peter Filzmoser, Kristel Joossens, and Christophe Croux, *Multiple group linear discriminant analysis: robustness and error rate*, Compstat 2006-Proceedings in Computational Statistics, Springer, 2006, pp. 521–532.

[7] John H Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence*, MIT press, 1992.

[8] Brian Kernan and John Geraghty, *A multi-objective genetic algorithm for extend*, Proceedings of the First Irish Workshop on Simulation in Manufacturing, Services AND Logistics, Limerick, Ireland, 2004.

[9] Br R Kowalski and MB Seasholtz, *Recent developments in multivariate calibration*, Journal of Chemometrics **5** (1991), no. 3, 129–145.

[10] W.J. Krzanowski, P. Jonathan, W.V. McCarthy, and M.R. Thomas, *Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data*, Applied Statistics **44** (1995), 101–115.

[11] Barry K Lavine, CE Davidson, Anthony J Moores, and PR Griffiths, *Raman spectroscopy and genetic algorithms for the classification of wood types*, Applied Spectroscopy **55** (2001), no. 8, 960–966.

[12] Tao Li, Shenghuo Zhu, and Mitsunori Ogihara, *Using discriminant analysis for multi-class classification: an experimental investigation*, Knowledge AND information systems **10** (2006), no. 4, 453–472.

[13] Howard Mark, *Chemometrics in near-infrared spectroscopy*, Analytica chimica acta **223** (1989), 75–93.

[14] Robert R. Meglen, *Chemometrics: Its role in chemistry and measurement sciences*, Chemometrics AND Intelligent Laboratory Systems **3** (1988), no. 1, 17 – 29, Proceedings of the Workshop in Chemometrics, Sponsored by the Environmental Protection Agency.

[15] Melanie Mitchell, *An introduction to genetic algorithm. massachussetts*, 1997.

[16] A. Mkhadri, G. Celeux, and A. Nasroallah, *Regularization in discriminant analysis: an overview*, Computational Statistics AND Data Analysis **23** (1997), no. 3, 403–423.

[17] Tormod Næs and Bjørn-Helge Mevik, *Understanding the collinearity problem in regression and discriminant analysis*, Journal of Chemometrics **15** (2001), no. 4, 413–426.

[18] Márcio José Coelho Pontes, Claudete Fernandes Pereira, Maria Fernanda Pimentel, Fernanda Vera Cruz Vasconcelos, and Alinne Girlaine Brito Silva, *Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification*, Talanta **85** (2011), no. 4, 2159–2165.

[19] Fabrice Rossi, Damien François, Vincent Wertz, Marc Meurens, and Michel Verleysen, *Fast selection of spectral variables with b-spline compression*, Chemometrics AND Intelligent Laboratory Systems **86** (2007), no. 2, 208–218.

[20] Carolina Santos Silva, Flávia de Souza Lins Borba, Maria Fernanda Pimentel, Marcio José Coelho Pontes, Ricardo Saldanha Honorato, and Celio Pasquini, *Classification of blue pen ink using infrared spectroscopy and linear discriminant analysis*, Microchemical Journal (2012).

[21] Urijatan Teixeira Carvalho Polari Souto, Márcio José Coelho Pontes, Edvan Cirino Silva, Roberto Kawakami Harrop Galv ao, Mário César Ugulino Araújo, Fátima Aparecida Castriani Sanches, Francisco António Silva Cunha, and Maria Socorro Ribeiro Oliveira, *Uvvis spectrometric classification of coffees by spalda*, Food Chemistry **119** (2010), no. 1, 368 – 371.

[22] Nidamarthi Srinivas and Kalyanmoy Deb, *Muiltiobjective optimization using nondominated sorting in genetic algorithms*, Evolutionary computation **2** (1994), no. 3, 221–248.

[23] Zou Xiaobo, Zhao Jiewen, Malcolm JW Povey, Mel Holmes, and Mao Hanpin, *Variables selection methods in near-infrared spectroscopy*, Analytica chimica acta **667** (2010), no. 1, 14–32.

[24] Jieping Ye, *Least squares linear discriminant analysis*, Proceedings of the 24th international conference on Machine learning (New York, NY, USA), ICML '07, ACM, 2007, pp. 1087–1093.